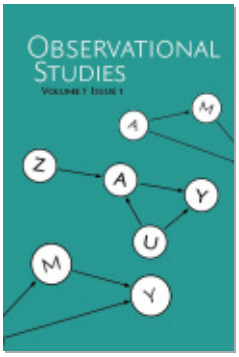
One Modern Culture of Statistics: Comments on Statistical Modeling: The Two Cultures (Breiman, 2001b)

Originally published in:
Observational Studies 7(1), https://doi.org/10.1353/obs.2021.0020

Funding acknowledgement:
786461 - Statistics, Prediction and Causality for Large-Scale Data (EC)

One Modern Culture of Statistics: Comments on Statistical Modeling: The Two Cultures (Breiman, 2001b)



Peter Bühlmann

Observational Studies, Volume 7, Issue 1, 2021, pp. 33-40 (Article)

Published by University of Pennsylvania Press
DOI: *https://doi.org/10.1353/obs.2021.0020*

Observational Studies Vol. 7.1 (2021): Published July 2021

One modern culture of statistics. Comments on Statistical Modeling: The Two Cultures (Breiman, 2001b)

Peter Bühlmann **buhlmann@stat.math.ethz.ch**
Seminar for Statistics
ETH Zürich
CH-8092 Zürich, Switzerland

Abstract

We comment on Leo Breiman's "Statistical Modeling: The Two Cultures" paper. We provide some thoughts on prediction from a broader perspective and argue that "aiming for one modern culture" is a highly embracing attempt for addressing key problems in data and information sciences.

Keywords: Causality, Distributional Robustness, Domain Adaptation, Generalization, Prediction, Statistical Machine Learning.

Introduction

Leo Breiman's "the two cultures" paper (Breiman, 2001b), his outstanding and highly original contributions in the "early" times of machine learning, and Leo as a scientist have all very much influenced my own thinking. His impact includes the creation of CART (Breiman et al., 1984), Bagging (Breiman, 1996), fundamental understanding of Boosting (Breiman, 1999) and further development of tree ensemble schemes (Amit and Geman, 1997) resulting in Random Forests (Breiman, 2001a). It was a time when prediction performance in the "classical regime" (e.g. for approximately i.i.d. data, see Section) has been significantly improved by new algorithms.

"The two cultures" paper (Breiman, 2001b) should perhaps be seen within this context. Over the last 20 years, many things have happened. In particular, the success of deep learning (Krizhevsky et al., 2012; LeCun et al., 2015) is shining into many areas. It has started largely about prediction in the "classical regime" (see Section), but then was gradually developed further to more reliable predictions in more realistic settings, e.g. with adversarial learning to improve robustness (Goodfellow et al., 2014) or with domain adaptation for better generalization (Pan et al., 2010). Almost every field in science and engineering uses deep learning or also Breiman's Random Forests for prediction purposes and associated variable importance; and a large mathematically oriented community contributes nowadays towards better understanding of the black box (the algorithmic model in Leo Breiman's terminology).

Breiman's thought-provoking distinction between the "two cultures", that is, the difference between a prediction and an interpretation ("data model", in Breiman's terminology) view-point can nowadays be better understood, as I will elaborate in Section . I should emphasize though that already 20 years ago, in response to the "two cultures paper", the

comments by David Cox (Cox, 2001) and Brad Efron (Efron, 2001) have pointed to some of these points.

A broader perspective on prediction

Leo Breiman argued in favor of prediction. Indeed, prediction can provide highly interesting answers to many problems which are often beyond the "classical regime". We describe some of the settings next.

The "classical regime". This is the scenario where the data are i.i.d. or stationary realizations of a single data generating probability distribution. Particularly in the former case, standard cross-validation schemes can be used for measuring empirical performance of prediction: something that has been done 20 years ago to demonstrate the empirical success of new algorithms. This is perhaps the setting which Leo Breiman had in mind: in fact his pictures on the first page (p.199) are re-displayed again in the current Figure 1 which contrasts the "classical regime" with more complicated scenarios.

External validity of predictions. The "classical regime" may not be realistic and empirical standard cross-validation is the wrong technique if the test data comes from a different distribution than the training sample. Already David Cox pointed out in his comment on "the two cultures" paper:

> The success of a theory is best judged from its ability to predict in new contexts (Cox, 2001).

It is often not sufficient to simply run cross-validation. Efron (2020) points to the difficulties with "concept drift" and provides an empirical example illustrating the problems with cross-validation. In many applications, one wants to generalize to new sub-populations which have never been observed during training of methods or algorithms. For example, transferability of algorithms (or "AI systems") for surveillance in intensive care units to a new hospital is often a difficult practical problem. The new data to be predicted may be very different from the training set and standard prediction tools will fail. Themes which try to address this issue include domain adaptation (Pan et al., 2010) or making prediction algorithms distributionally robust (Sinha and Namkoong, 2017; Gao et al., 2017).

Causal effects. At another pole of the spectrum is the prediction of what happens when an external intervention or manipulation is done or a new policy is implemented. The effect (e.g. the derivative $\frac{\partial}{\partial x}$ of the expected value when intervening with an intervention value x) of such an intervention on a response variable of interest is the total causal effect. Methods and algorithms which would perform well for such questions would be extremely powerful. This goal is very ambitious in absence of randomized interventions (in randomized studies): and thanks to some advances in causality and causal inference, we know that there are severe limitations when having only access to observational data. But we want to highlight here that standard prediction algorithms, even if they were perfect with minimal Bayes error in the "classical regime" (mentioned above), cannot directly provide answers for the prediction of interventions. Sometimes though, this fundamental fact is forgotten and very naive conclusions are drawn from data. To illustrate the point we quote Brad Efron who wrote 20 years in response to "the two cultures" paper:

> Estimation and testing are a form of prediction: "In our sample of 20 patients drug A outperformed drug B; would this still be true if we went on to test all possible patients?" ...(Peter Gregory) undertook his study for prediction purposes, but also to better understand the medical basis of hepatitis. Most statistical surveys have the identification of causal factors as their ultimate goal (Efron, 2001).

We contrast the pictures on the first page from Breiman (2001b) with a perhaps more realistic picture in Figure 1.

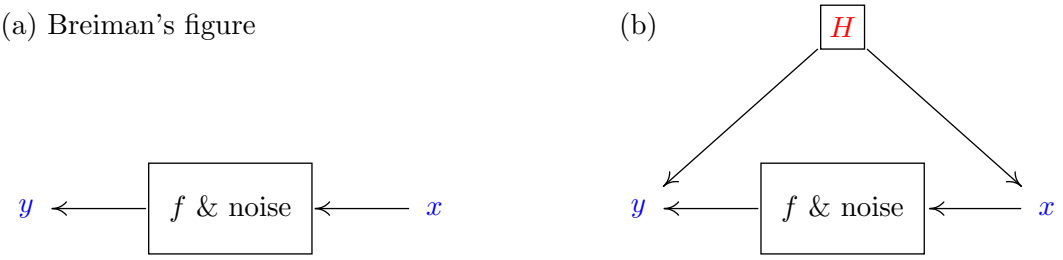


Figure 1: Left panel (a): Breiman's pictures (symbolized by one figure) for (mostly) the "classical regime". Right panel (b): Modification of (a), emphasizing the issue that hidden variables, e.g., encoding different environments or subpopulations or generating perturbations to the "true" system. The task is still to predict y from x, but ignoring H may result in very poor performance.

Connections between the settings

Perhaps surprising at first sight, the settings described above are related. The causal prediction which is based on the causal covariates only can be represented as an optimal solution to a distributionally robust optimization (Peters et al., 2016; Meinshausen, 2018; Rothenhäusler et al., 2018; Bühlmann, 2020). Informally, one can think of the following worst-case risk optimization for regression with the L_2-error:

$$f^*_{\text{causal}} = \text{argmin}_{f \in \mathcal{F}} \max_{P \in \mathcal{P}} \mathbb{E}_P[(Y - f(X))^2], \tag{1}$$

where $\mathcal{P}$ is a suitable set of distributions modeling potential arbitrarily strong perturbations, interventions or heterogeneous data generating mechanisms, $\mathcal{F}$ is a suitable function class, and X denotes the p-dimensional covariates and Y a univariate response of interest. The support of $f^*_{\text{causal}}(\cdot)$ are the causal covariates for Y.

Furthermore, distributionally robust solutions can be typically represented as penalized versions of the prediction algorithms for the "classical regime" (Sinha and Namkoong, 2017, cf.). The worst case risk optimization problem in (1) can be (typically) be reformulated in dual form:

$$\text{argmin}_{f \in \mathcal{F}} \mathbb{E}[(Y - f(X))^2] + \gamma \text{pen}_{\text{causal}}(f), \tag{2}$$

for some appropriate penalty function $\text{pen}_{\text{causal}}(\cdot)$ and where $0 \leq \gamma \leq \infty$. This penalty term is tailored towards causality and hence denoted with the sub-index "causal": it builds on the idea of encouraging a certain distributional invariance of residuals $Y - f(X)$ over a class of distributions. In practice one always replaces the expectation by the empirical mean. We emphasize that we are penalizing here the population version (2): it is not a complexity penalization due to finite sample size, but it rather serves as a step for robustification. The interesting fact, proven under certain assumptions on $\mathcal{P}$ and for linear function classes $\mathcal{F}$ (Rothenhäusler et al., 2018), is the following:

$$\begin{aligned} &\gamma = \infty \text{ leads to the causal solution } f^*_{\text{causal}}(\cdot), \\ &0 < \gamma < \infty \text{ leads to some distributional robustness, increasing as } \gamma \to \infty. \end{aligned} \tag{3}$$

We thus conclude from this that the causal solution $f^*_{\text{causal}}(\cdot)$ is maximally robust (under some assumptions). If one would know the causal components of X to Y, a highly ambitious goal when having observational data only, one could use regression (or classification) based on the causal variables which would lead to robust predictions. Or in other words: the causal solution leads to robust prediction in new contexts, something which is rather "obvious" to the experts. The connection indicated by (1)-(3) provides a quantitative relation and understanding. The causal solution is "more simple than the classical prediction machine" as it puts more restrictions in the causal regularization (2); and distributional robustness can be obtained by using a "an intermediate amount of simplicity". Equations (2) and (3) provide a justification for the folklore statement that "simple methods are more robust and may generalize better in difficult problems".

The availability of deep neural networks or Leo Breiman's Random Forests can be of great value as a "workhorse". But in my view, it is often good to regularize them back to gain more robustness, invariance across different contexts or simplicity, as outlined in (2). And perhaps, a more simple statistical model fit may then perform comparably or is even more powerful than a complicated plain vanilla prediction algorithm. Another form of regularizing "towards causality" is with a (approximately "true") physical model, e.g. in the form of differential equations, combined with statistical estimation techniques (Ramsay, 2000). As Cox (2001) wrote: a method, algorithm or theory should be "judged from its ability to predict in new contexts".

Two, many or one culture?

John Tukey in his famous 1962 paper "The future of data analysis" (Tukey, 1962) has opened a new path for data analysis and information extraction from data. Tukey writes in the final section:

> What of the future? ... That remains to us, to our willingness to take up the rocky road of real problems (Tukey, 1962).

He argued for embracing a change of culture and expanding the horizon of mainstream statistics. 15 years later, adopting the call for a change of culture, Leo Breiman organized in 1977 a session on "Large and Complex DataSets" at an ASA- and IMS-sponsored conference in Dallas. Jerry Friedman wrote in 1998 about "Data Mining and Statistics: What's the

connection?" (Friedman, 1998) and argued strongly for a change and openness to other directions:

> Perhaps more than at any time in the past, Statistics is at a crossroads; we can decide to accommodate or resist change (Friedman, 1998).

There were others who expressed similar views. And then, Leo Breiman published in 2001 his influential "the two cultures" article, again with the intention to expand the horizon of the statistics community; as opposed to divide the community into two separate areas. The paper is perhaps written in an asymmetric form, namely addressing the statistical community (and using the outlet of Statistical Science) to embrace new developments outside statistics. As indicated above, there have been earlier and also later papers which have emphasized this point of absorbing new ideas from outside statistics, e.g. machine learning and other application oriented fields. In the other direction, Tom Mitchell writes in 2006

established within statistics	at the "periphery" of statistics
Non-, Semiparametric and High-dimensional Estimation	Deep neural networks, Double Machine Learning Ensemble Methods, Kernel machines
Causal Inference	Explainable Algorithms Reinforcement Learning
Statistical Robustness	Adversarial Learning
Design of Experiments	Active Learning
Statistical Learning Theory	Mathematical Foundations for Algorithms and Deep Learning
Computational Statistics Statistical Computing	(large-scale) Optimization
Statistical Modeling Applied Statistics, Validation	Machine Learning Applications, Veridical Data Science (Yu and Kumbier, 2020)
Replicability Uncertainty Quantification	– –
CART, Bagging, Random Forests	*CART, Bagging, Random Forests*

Table 1: Related topics: established in statistics (left) and still at the periphery of statistics (right). The dashes "–" indicate that these topics are mostly built on established ideas from statistics, Some important contributions by Leo Breiman are displayed in emphasis mode: thanks to his vision and excellence, they are established within and outside statistics.

in his book on "The discipline of machine learning" (Mitchell, 2006):

> Machine learning will help reshape the field of statistics, ... Of course both computer science and statistics will also help shape machine learning as they

progress and provide new ideas to change the way we view learning (Mitchell, 2006).

David Donoho's "50 years of data science", written in 2017 (Donoho, 2017), provides an excellent view on the broad field of data science and how different communities have contributed to the field. The "50 years" refer to John Tukey's famous 1962 paper "The future of data analysis". Tukey branched out and initiated a new direction, and so did Breiman also with his "the two cultures paper". Why two cultures? Breiman pointed out that statistics should embrace an additional culture of "prediction and algorithmic modeling". I think though that other facets, perspectives and approaches should co-exist: indeed, statistics has always been close to essentially any field where data analysis is done. We can argue for many cultures, or for one culture as discussed next. I am much more inclined to the idea of one culture. The term "one culture" has been used by Yu and Barter (2020). It is about "one (broad) culture" in the field of statistics, building on well-established and proven methodology as well as using new tools and exploiting developments outside statistics. I am exemplifying with some selective examples in Table 1: the distinction between "established" and "peripheral" is often blurred though. The established principles are still very useful; and they can serve as a "benchmark", not only for prediction, but to also assess the usefulness and complementary nature of new developments which might be inspired by other fields. Combining "established" and "peripheral" would imply that peripheral (from the viewpoint of statistics) developments move also to the center of statistics, and it will eventually make the discipline of statistics a powerful and highly credible key-player in the data and information sciences.

Acknowledgments

We thank Armeen Taeb for some useful comments. This research is supported by the European Research Council under the Grant Agreement No 786461 (CausalStats - ERC-2017-ADG).

References

Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9:1545–1588, 1997.

L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

L. Breiman. Prediction games & arcing algorithms. *Neural Computation*, 11:1493–1517, 1999.

L. Breiman. Random Forests. *Machine Learning*, 45:5–32, 2001a.

L. Breiman. Statistical modeling: The two cultures (with discussion). *Statistical Science*, 16:199–231, 2001b.

L. Breiman, J. Friedman, C. Stone, and R. Olshen. *Classification and Regression Trees*. CRC press, 1984.

P. Bühlmann. Invariance, causality and robustness (with discussion). *Statistical Science*, 35:404–426, 2020.

D. Cox. Comment on "Statistical modeling: The two cultures". *Statistical Science*, 16: 216–218, 2001.

D. Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26:745–766, 2017.

B. Efron. Comment on "Statistical modeling: The two cultures". *Statistical Science*, 16: 218–219, 2001.

B. Efron. Prediction, estimation, and attribution. *Journal of the American Statistical Association*, 115:636–655, 2020.

J. Friedman. Data mining and statistics: What's the connection? *Computing science and statistics*, 29:3–9, 1998.

R. Gao, X. Chen, and A. Kleywegt. Wasserstein distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050*, 2017.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks, 2014. arXiv preprint arXiv:1406.2661.

A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015.

N. Meinshausen. Causality from a distributional robustness point of view. In *2018 IEEE Data Science Workshop (DSW)*, pages 6–10. IEEE, 2018.

T. Mitchell. The discipline of machine learning, 2006. Carnegie Mellon University, School of Computer Science, Machine Learning Department. Available at `http://ra.adm.cs.cmu.edu/anon/usr0/ftp/anon/ml/CMU-ML-06-108.pdf`.

S.J. Pan, I. Tsang, J. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2010.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference using invariant prediction: identification and confidence interval (with discussion). *J. Royal Statistical Society, Series B*, 78:947–1012, 2016.

J. Ramsay. Differential equation models for statistical functions. *Canadian Journal of Statistics*, 28:225–240, 2000.

D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. Anchor regression: heterogeneous data meets causality. *To appear in the J. Royal Statistical Society (Series B); preprint arXiv:1801.06229*, 2018.

A. Sinha and J. Namkoong, H.and Duchi. Certifiable distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017. Presented at Sixth International Conference on Learning Representations (ICLR 2018).

J. Tukey. The future of data analysis. *The Annals of Mathematical Statistics*, 33:1–67, 1962.

B. Yu and R. Barter. The data science process: one culture. *Journal of the American Statistical Association*, 115:672–674, 2020.

B. Yu and K. Kumbier. Veridical data science. *Proceedings of the National Academy of Sciences*, 117:3920–3929, 2020.